Project review

- Title: SDCI Net: Collaborative Research: An integrated study of datacenter networking and 100 GiGE wide-area networking in support of distributed scientific computing
- Participants: University of Virginia, University of New Hampshire, National Center for Atmospheric Research (NCAR)
- Date: Oct. 25, 2013

Supported by NSF grants: OCI-1127340, OCI-1127228, OCI-1127341
Questions on this slide set? Contact Malathi Veeraraghavan, mv5g@virginia.edu

Agenda

- 10:30-10:50: SDCI Project overview, MV, UVA
- 10:50-11:10: Zhengyang Liu, UVA (NDM paper)
- 11:10-11:20: Scott Tepsuporn, UVA (FT engineering)
- 11:20-11:30: Leiqing Cai, UVA (MPI work)
- 11:30-12:30: Bob Russell and Patrick MacArthur, UNH
- 12:30-1:00: Lunch/Discussion
- 1:00-1:40: John Dennis, NCAR
- 1:40-2:00: Zhengyang Liu, NCAR (MPI)
- 2:00-2:30: Feedback/Plans on NSF SDCI project

Acknowledgment

- UVA graduate and UG students on our project
  - Presenters + Zihao Wang, UG RA + Zhenzhen Yan, GRA
- NSF OCI, co-PIs and students
  - Kevin Thompson, Bob Russell, John Dennis, Patrick MacArthur, Fabrice Micero
- ESnet:
  - Chris Tracy, Brian Tierney, Inder Monga, Greg Bell, Jon Dugan, Andy Lake, Tareq Saif, and Eric Pouyoul
- UVA Stats department: Jianhou Zhou
- NERSC: Jason Hick
- SLAC: Yee-Ting Li and Wei Yang
- ANL: Raj Kettimuthu

Year 2 Accomplishments

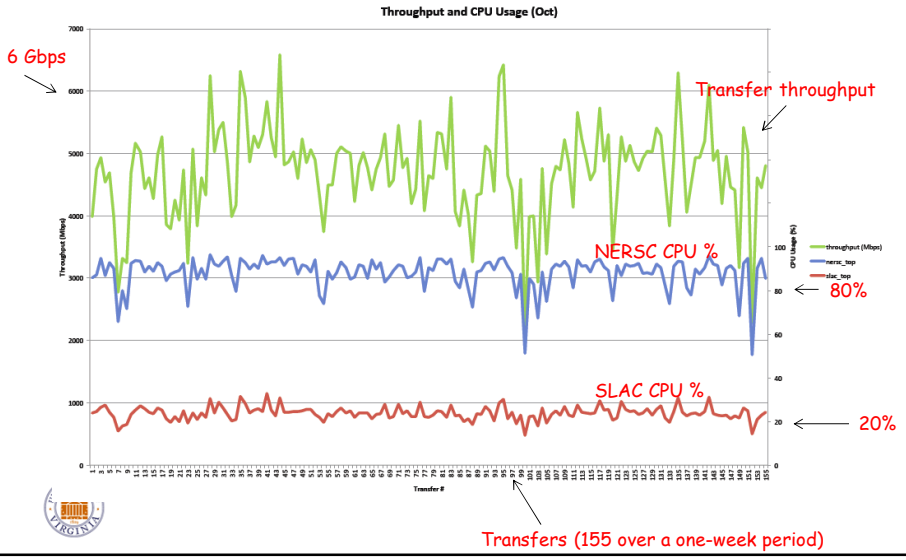
- Characterized sources of file-transfer throughput variance
  - Key contribution: method to create non-linear regression models that capture dependence of throughput on various factors
  - Accepted paper at NDM workshop, in assoc. with SC 2013
- QoS experiments showed limitations of policing VCs
  - published CTRQ 2013 paper; accepted journal version
- Design of a low throughput-variance file-transfer solution
- Datacenter networking (reduce MPI latency for CESM apps)
  - Developed a four-phase approach in an week-long meeting of team
  - Started implementation over the summer
  - MPI, RDMA, InfiniBand interaction work

NDM paper summary

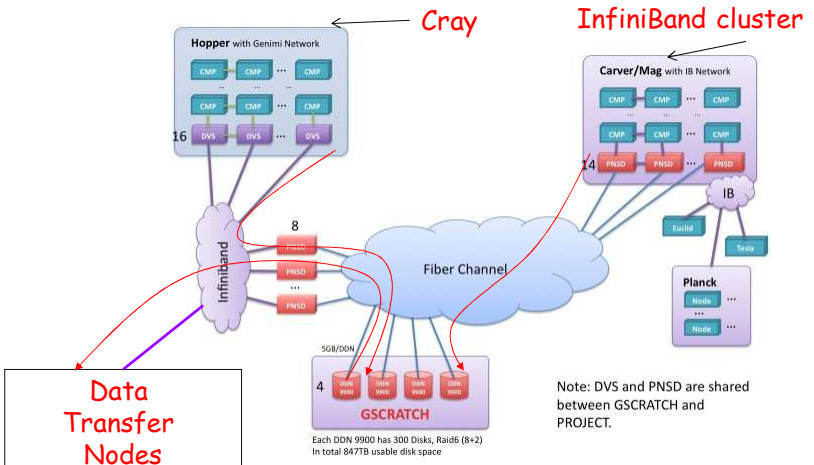
- Tests between production data-transfer nodes (DTNs) showed the following:
  - amount of CPU time assigned to a particular file-transfer process was a key determinant of mem2mem throughput (since interactive logins are allowed on DTNs)
  - measured 10^{-5} average packet loss rate across metro-area path with 10 routers, but contribution to throughput was -16% ⇒ virtual circuits can cut this loss rate
  - CPU time needed at two ends were highly correlated – if managed, need concomittant resource allocations
  - Disk I/O competition another significant factor
- Developed method for creating non-linear regression models for throughput

5

Transfer throughput dependence on CPU time



Example: NERSC systems (competition for disk access)



QoS experimental findings

- To use virtual circuits for file transfers:
  - when a VC is requested with a certain rate, policing is configured on ingress side of router interface to limit flow to that rate
  - out-of-profile packets were shunted to a scavenger service queue in Internet2/ESnet config.
    - experiments showed that this leads to TCP throughput drops because of out-of-sequence packets
  - WRED tried for out-of-profile packets, but with two high-speed flows, packet drops leads to worse throughput than w/o policing

For high-throughput on VCs

- Choice between
  - low-variance, potentially increased waiting time for a high-throughput VC
  - high-variance, lower waiting time for a high-throughput IP-routed path
    - elephants stomp over mice
- if a DTN pair can sustain high throughput, request a VC for the high rate, but wait time may be more
  - to lower waiting time, lower utilization

Engineering solution

- For low-variance, high-throughput transfers (part of workflow)
  1. run managed FT processes on DTNs (disable interactive; dedicated nodes for large dataset transfers)
  2. leverage Science DMZ to stage a local copy from shared filesystems
     - need a 2-phase cycle to alternate between unmanaged transfers from shared file systems and managed transfers across WAN
  3. callibrate required CPU times at two ends and VC rate (create nonlinear regression models – server dependent)
  4. schedule FT processes at two ends with PBS and schedule VC
- RoCE will reduce dependence on CPU time, but still need to schedule FT process on CPUs because of disk access competition
  - different nonlinear regression model for determining resource requirements

Datacenter networking metric: MPI latency

- Had an in-person meeting of whole team at UVA
- Generated a plan to execute 4 tasks
  - Experimental execution of MPI pingpong application and obtain measurements (using ibdump) to determine if competing traffic from other applications on Yellowstone cause variance in MPI latency
  - Run CESM apps on Yellowstone and obtain traces and ibdump output and analyze
  - RIB-type analysis of IB switch routing tables – centralized subnet manager controls routing table in InfiniBand network
  - Joint simulation of computation and communication

Plan for this year

- Prototype FT engineering solution with controlled low-variance, but high throughput (Scott Tepsuporn)
- Intra-datacenter networking
  - MPI latency "discovery" phase (Leiqing Cai)
  - IB routing schemes to develop engineering solution to address latency problem (Zihao Wang)
- FT application monitoring GUI

FT application monitoring GUI

- Plan for next year
  - Improve GridFTP performance GUI
  - Demo'ed last year
  - Pursuing adoption by ANL GridFTP group?
- Could potentially add to PerfSONAR

Agenda

- 10:30-10:50: SDCI Project overview, MV, UVA
- 10:50-11:10: Zhengyang Liu, UVA (NDM paper)
- 11:10-11:20: Scott Tepsuporn, UVA (FT engineering)
- 11:20-11:30: Leiqing Cai, UVA (MPI work)
- 11:30-12:30: Bob Russell and Patrick MacArthur, UNH
- 12:30-1:00: Lunch/Discussion
- 1:00-1:40: John Dennis, NCAR
- 1:40-2:00: Zhengyang Liu, NCAR (MPI)
- 2:00-2:30: Feedback/Plans on NSF SDCI project